PHONETIC CONSEQUENCES OF SPEECH DISFLUENCY

Elizabeth E. Shriberg

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA 94025
ees@speech.sri.com

ABSTRACT

Unlike read or laboratory speech, spontaneous speech contains high rates of disfluencies (e.g., repetitions, repairs, filled pauses). Such events reflect production problems frequently encountered in everyday conversation. Analyses of American English show that disfluency affects a variety of phonetic aspects of speech, including segment durations, intonation, voice quality, vowel quality, and coarticulation patterns. These effects provide clues about production processes, and can guide methods for disfluency processing in speech recognition applications.

1. INTRODUCTION

A clear difference between spontaneous speech and read or laboratory speech is that the former contains significant rates of disfluencies (e.g., filled pauses, repetitions, and repairs), such as

Filled pause	"uh", "um"
Repetition	"the the"
Repair	"any health cov- any health insurance"
False Start	"It's fir- I could get it where I work"

In laboratory or read speech, where content is given or highly constrained, minimal complex processing is required. But in everyday conversation, our messages are constructed on the fly. We must decide what to say, how to say it, and how to coordinate our interactions with others—all in real time. It is thus hardly surprising that we sometimes need to pause, or to repair our previous speech.

Rates of disfluency per word in spontaneous English speech vary from under 1% for constrained human-computer dialog, to 5-10% for natural conversations [12, 16, 4, 19]. There is also considerable variation across speaking environments in the relative rates of particular disfluency types [12, 19]. Such distributional differences reflect differences in cognitive demands, and in managing interaction in conversation [9, 2].

While considerable past work has focused on lexical properties of disfluency, recent years have seen increasing attention to the question of phonetic properties. An early suggestion by Hindle [7] proposed that disfluencies are marked by a special acoustic "edit signal" at interruption. Although inspection [1], as well as psycholinguistic experiments [11], has revealed no such specific signal, disfluency is nevertheless associated with a variety of phonetic characteristics that differentiate it from fluent speech.

The goal of this paper is to outline some of the main phonetic consequences of disfluency. As we will see, such effects can provide a window onto production processes that a lexical or word-level analysis often obscures. They can also guide development of improved models for disfluency processing in speech applications.

2. THE STRUCTURE OF DISFLUENCIES

The majority of disfluencies that occur in spontaneous speech can be analyzed as having the following three-region surface structure (terms adapted from Levelt [8]):

(Prior context)	Repar- andum	Editing Phase	Repair	(Continuation)
	.	um		we're fine
it's	.	uh		after five
have	the .		the	tools
to	res- .		relax	at home
all	this .		this	paper
	it's .			did you?

The first region of the disfluency is the *reparandum*, or material that will later be replaced. The end of this region corresponds to the *interruption point* (marked with a ".") or the location at which there is a departure from fluency. By this point, the speaker has detected some problem, and according to a "Main Interruption Rule" halts the production process [8].) The editing phase consists of the region from the interruption point to the onset of the repair. This region may be empty, contain a silent pause, or contain editing phrases or filled pauses ("I mean", "um", "uh"). The term "editing" is not intended to imply detection of error; pausing can occur for reasons not involving error. Finally, we have the repair region, which typically reflects the resumption of fluency. (We will assume here that the repair is not itself followed by another self-interruption. If it is, the disfluency is "complex" [19].) These regions are contiguous, and removal of the first two (reparandum and editing phase) yields a lexically "fluent" version.

As shown, we can analyze all of our disfluency types this way. A disfluency may contain material only in the editing phase, such as a filled pause. Or it may contain only repeated words in the reparandum and repair. Note that for repeats such as "the the", this structure predicts that it is the first instance, and not the repeated one, that is most likely to be aberrant, a prediction we will see later evidence for based on phonetic features. Editing terms can combine with different types of disfluency (e.g., "the uh the"; "res- i mean relax").

We will organize our overview of phonetic consequences by moving through these three regions left to right, discussing the effects in each. As we will see, most of the properties are in the reparandum and editing phase, but certain effects can also be seen in the repair.

3. EFFECTS IN THE REPARANDUM

Although at a lexical level of representation the reparandum is removed in full to arrive at a fluent lexical version, it is not until

1. REPORT DATE **1999**	2. REPORT TYPE	3. DATES COVERED **00-00-1999 to 00-00-1999**
4. TITLE AND SUBTITLE **Phonetic Consequences of Speech Disfluency**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **SRI International,Speech Technology and Research Laboratory,Menlo Park,CA,94025**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT **see report**

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES **4**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

the speaker notices trouble that we should expect to see phonetic manifestations. Indeed this is what we find. Phonetic effects in the reparanda of disfluencies are most prevalent at or around the interruption point.

3.1. Duration Patterns

One of the most pervasive effects of disfluency is a lengthening of rhymes or syllables immediately preceding the interruption point. Effects can at times be seen to extend further back, into full words or phrases. As an example we will look at one-word repetitions such as "the the" in the Switchboard corpus of human-human telephone conversations [6]. To assess lengthening, we compare the durations of each instance to the duration of "the" in fluent contexts. Results are shown in Figure 1; they represent data from a single speaker. As can be seen, the reparandum (Rep1) is length-
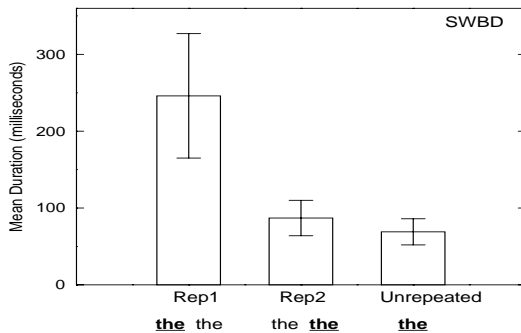

Figure 1. Duration of Words in Repetitions and Fluent Counterparts. R1=1st instance (reparandum), R2=2nd instance (repair).

ened considerably relative to its expected duration in fluent speech, whereas the repair (Rep2) has about the same duration as the fluent counterpart. This suggests that in repetitions, speakers are drawing out the reparandum much like they might a filled pause. However not all repetitions show this pattern. A more detailed study shows that there are at least three main types of repeats when classified based on prosodic properties, suggesting at least three different underlying states of the speaker when repeating [17]. The pattern depicted in Figure 1 however corresponds to the most common case. Durational lengthening in the reparandum is seen for other disfluency types as well, and is one way speakers can pause without ceasing phonation [3].

3.2. Intonation

Interestingly, when speakers modify duration, they tend to do so in a way that preserves intonation patterns and local pitch range relationships. For example while the reparandum in a repetition is often extended in duration, it typically shows a pitch contour similar to that of its following counterpart in the repetition, but just stretched out over more time—as shown in Figure 2 (pitch tracks are indicated for only the words in the repetition):
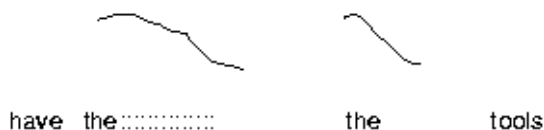

Figure 2. Pitch of Repeated Words.

3.3. Word Cutoffs and Laryngealization

In read or laboratory speech, we expect words to be completed, but this is not the case in spontaneous speech. Speakers halt production soon after noticing trouble [8], without concern for word boundaries. In a corpus of human-computer dialog on air travel planning (ATIS; [13]) nearly 60% of disfluencies contained word cutoffs; rates in two human-human corpora were about 20-25% [19]. The difference is largely due to the higher relative rate of error repairs in human-computer dialog. Errors are not higher overall in such corpora, but because non-error hesitations (filled pauses and repetitions) are suppressed in human-computer dialog with a push-to-talk mechanism for speech input, errors make up a larger proportion of total disfluencies.

Various researchers have described cutoffs as abrupt, showing some form of laryngealization [1, 14, 11]. In an analysis of cutoffs in the ATIS data conducted by Madelaine Plauché, we found that a typical form of laryngealization in such cases corresponds to creaky voice on the last 20-50 ms of the cut off words. However, it is not the case that all cutoffs are sudden, or that word cutoffs always correspond to errors. On the contrary, the highest rate of cutoffs found in the ATIS corpus was on simple repetitions. Here the rate was over 70% of repeats, whereas rates for repairs of error were under 50%. And some cutoffs could be extended in duration, more indicative of hesitation than of sudden detection of error.

Cut off words present a problem for automatic speech recognition since partial-word pronunciations are not present in the dictionary. Although one could add all possible initial phone sequences of a word as possible pronunciations, such an approach would create a proliferation of pronunciations that would only hurt performance by increasing confusability. A possible solution is to constrain fragments to be recognized only as parts of closely following words.

3.4. Coarticulation

Another consequence of disfluency is a change in surface coarticulation patterns. In the production of words in fluent speech, articulators generally move toward the articulator positions for the onset of the next word. But in disfluencies, this proximal relationship of coarticulation to actual output word sequence cannot be assumed. Coarticulation is governed by the next word in the speaker's phonetic *plan* at the time the word in question is uttered—not by the word sequence that is ultimately produced. In fluent speech, the plan and the final output are consistent, but in disfluencies, following lexical content may be temporarily unavailable, or the plan can change on the fly.

We looked at this question in a study of single-word repeats of "the" and "I". Note that only the place of articulation can safely be determined for transitions, although there are some cases where the manner is clear. We will classify cases based on consistency with a surface word, although of course we cannot know for sure whether some other word was intended. Below are results with illustrative examples; transitions are marked in parentheses, using standard orthography.

Transition	Frequency	Example
(a) NONE	722 (88%)	the . the dog
(b) to word after repeat	71 (9%)	the(d) . the dog
(c) to different word	19 (2%)	the(d) . the cat
(d) to repeat itself	3 (.3%)	the(th) . the dog

As shown, most cases of repeats have no detectable final transition. This is different from what is expected in fluent connected speech; here most cases contained a pause at interruption. For

speech recognition models, we may thus want to turn off cross-word modeling at repetition boundaries, or more generally at the interruption point of disfluencies.

The next two cases are also quite interesting, because they show coarticulation that is inconsistent with the following surface word. Case (b), which represents the majority of cases with coarticulation, shows that sometimes disfluency effects can be seen earlier than the location of the element causing trouble. From the transition we can infer that the speaker committed to the word directly after the repetition but stalls earlier, perhaps to keep syntactic or prosodic units intact. Case (c) is almost certainly a covert repair, where some word other than "cat" was caught before it was uttered, and repaired. Case (d) is standard in terms of having a transition consistent with the actual following word, but notice that the following word is the repeat itself. This suggests that in some cases, speakers must be planning to repeat while they are still producing the first instance of the word. As with case (a), cases (b) and (c) also pose problems for cross-word modeling in speech recognition. This time, the problem is that there is acoustic evidence for a segment at the end of the reparandum that is inconsistent with recognizer models constrained to model pronunciation only across contiguous surface words.

3.5. Vowel Quality

Disfluency is also associated with alterations in vowel quality. A special case is the word "the", which has an alternate pronunciation, [dh iy], before vowel-initial words in many dialects of American English. This alternate is also more likely in the reparandum of repetitions [5]. Other words without such variants, but with citation forms that differ from their pronunciation in connected speech, show a similar behavior (although it is not clear whether they reflect the same phenomenon). For example, "a" and "to" are much more likely to be pronounced as [ey] and [t uw] in the reparandum of disfluencies than elsewhere. It is not clear whether such forms are produced as "signals" to listeners, or whether they reflect a modification related to other acoustic properties such as durational lengthening and following pauses. However it is clear that speakers choose the alternate form before uttering the word, because vowel quality never shifts within the word itself.

4. EFFECTS IN THE EDITING PHASE

4.1. Unfilled Pauses

Under Levelt's framework of speech production [9], self-interruption is associated with a halting of the speech production process at all levels. Therefore, some minimum time is needed after the speech is cut off in order to plan the repair. Disfluency is thus often indicated by unfilled pauses in the editing phase. For automatic speech processing of disfluencies, these pauses have proven to be very useful. Work using decision trees to model acoustic features finds that pauses are among the best cues to disfluency detection [20, 21], because they are robustly extracted and ensure high recall.

4.2. Filled Pause Duration

In English, the vowel in the filled pauses "um" and "uh" is typically close to schwa; however, it can also carry stress, or occur further back and lower in the vowel space. In automatic speech recognition, Filled pauses are sometimes misrecognized as "a" or as parts of other words containing the relevant vowels. But filled pauses differ dramatically from these other instances in duration. To illustrate, durations for the vocalic portion of 700 filled pauses and for 40,000 instances of the same vowels elsewhere, including in the determiner "a", were obtained from recognizer forced alignments using the ATIS corpus. Results are shown in Figure 3.
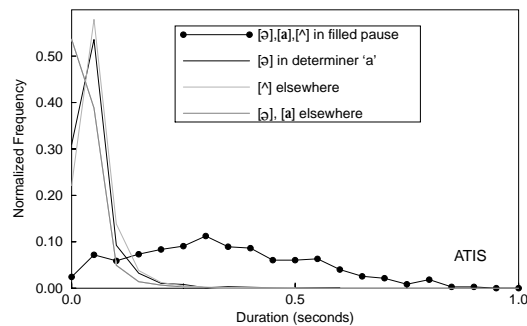

Figure 3. Duration of Vowels in Filled Pauses and Elsewhere.

As shown, vowels in filled pauses have much longer durations than the same vowels in fluent contexts. Duration, then, is a simple cue that could be used by speech recognition systems in discriminating vowels in filled pauses from the same vowels elsewhere. It is also important to treat such durations separately in duration modeling for other purposes, so as not to skew the distributions for these vowels.

4.3. Filled Pause Intonation

Filled pauses have been shown to be low in F0, and display a gradual, roughly linear pitch or fundamental frequency (F0) fall [15]. In addition, the F0 of filled pauses occurring within a clause was found to be related to the F0 of the surrounding speech [18]. Figure 4 shows F0 values for the onset and offset of a filled pause, and the preceding and following F0 peaks. Lines connect points for a specific filled pause. The four F0 measurements are plotted at equally spaced intervals; therefore the actual temporal intervals between these points (which varied greatly) are not represented in the figure. The solid heavy line indicates the speaker's estimated "baseline" F0, as estimated by measuring F0 at the end of sentence-final F0 falls.
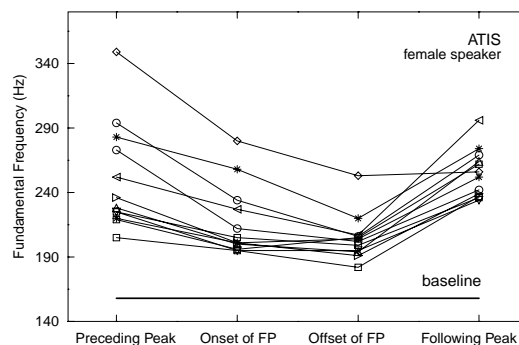

Figure 4. F0 of Filled Pauses and Surrounding Peaks.

What is striking here is that the F0 of filled pauses falls about halfway between the preceding peak value and the speaker baseline. In fact, F0 values in the study were well predicted by a simple additive-multiplicative model based on these values. These relationships held despite considerable differences in time intervals between the four measured values plotted at regular intervals as in Figure 4. These findings suggest that for filled pauses, similar to what we saw earlier for repetitions, speakers may preserve intonational relationships under changes in duration necessitated by the need to pause.

5. EFFECTS IN THE REPAIR

As said earlier, most consequences of disfluency are located in the reparandum and editing phase, since the repair region constitutes the onset of fluency. An exception, however, is that certain types of repair can show effects of having made a change in content, in the form of contrastive emphasis on the repairing element.

Levelt and Cutler [10] looked at prosodic marking, or an increase in F0, duration, or amplitude, in the repair region of disfluencies from a pattern description task. They found that marking occurred for roughly half of the repairs involving error, and for only about 20% of the repairs involving mere elaboration. This suggests that it may be more important to call attention to outright error than to inappropriateness. Such marking also illustrates that we cannot simply remove the reparandum and editing phase, leaving a perfectly fluent repair. All three regions are still in the discourse record; the prosodic contrast in the repair is produced with respect to the earlier mention in the reparandum.

6. SUMMARY AND CONCLUSION

Disfluencies are rare in laboratory speech, but occur with considerable frequency in everyday communication. Most disfluencies can be analyzed as having a three-region structure, in which the first two regions are removed to yield a fluent version of the utterance. Disfluency affects a variety of phonetic aspects of speech, mainly in the two regions that are removed. The effects include changes in segment durations, intonation, word completion, voice quality, vowel quality, and coarticulation patterns. These effects provide insights into the mechanisms underlying the production of spontaneous speech in conditions characteristic of the real world. In addition they provide information that can be used to better model disfluencies in automatic speech recognition applications.

REFERENCES

[1] Bear, J., Dowding, J., and Shriberg, E. E. 1992. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, University of Delaware, Newark, Delaware.

[2] Clark, H. 1996. *Using Language*. Cambridge: Cambridge University Press.

[3] Duez, D. 1993. Acoustic correlates of subjective pauses. *Journal of Psychological Research*, 22(1), 21–39.

[4] Fox Tree, J. E. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34, 709–738.

[5] Fox Tree, J. E. and Clark, H. H. 1997. Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, 62, 151–167.

[6] Godfrey, J., Holliman, E., and McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 517–520.

[7] Hindle, D. 1983. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pp. 123–128.

[8] Levelt, W. J. M. 1983. Monitoring and self-repair in speech. *Cognition*, 14, 41–104.

[9] Levelt, W. J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

[10] Levelt, W. J. M. and Cutler, A. 1983. Prosodic marking in speech repair. *Journal of Semantics*, 2(2), 205–217.

[11] Lickley, R. J. 1994. *Detecting Disfluency in Spontaneous Speech*. Ph.D. thesis, University of Edinburgh.

[12] Maclay, H. and Osgood, C. E. 1959. Hesitation phenomena in spontaneous English speech. *Word*, 15, 19–44.

[13] MADCOW. 1992. Multi-site data collection for a spoken language corpus. In *Proceedings of the 5th DARPA Speech and Natural Language Workshop*. Morgan Kaufmann.

[14] Nakatani, C. H. and Hirschberg, J. 1994. A corpus-based study of repair cues in spontaneous speech. *Journal Acoustical Society of America*, 95(3), 1603–1616.

[15] O'Shaughnessy, D. 1992. Recognition of hesitations in spontaneous speech. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, I, 521–524.

[16] Oviatt, S. L. 1995. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9, 19–35.

[17] Plauché, M. and Shriberg, E. 1999. Data-driven subclassification of disfluent repetitions based on prosodic features. In *Proceedings of the International Congress of Phonetic Sciences*, San Francisco.

[18] Shriberg, E. E. and Lickley, R. J. 1993. Intonation of clause-internal filled pauses. *Phonetica*, 50, 172–179.

[19] Shriberg, E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California, Berkeley, CA.

[20] Shriberg, E., Bates, R., and Stolcke, A. 1997. A prosody-only decision-tree model for disfluency detection. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, vol. 5, pp. 2383–2386.

[21] Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plauché, M., Tür, G., and Lu, Y. 1998. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proceedings of the International Conference on Spoken Language Processing*, vol. 5, pp. 2247–2250.